DESCRIPTION

The present invention relates to the field of taxonomic analysis of biological samples, based on the use of the DNA sequence of a highly evolutionarily conserved protein.

PRIOR ART

The taxonomic analysis of samples is applicable to a broad spectrum of industries and activities. The three main areas of application of taxonomic analysis are:

Analysis of foodstuffs and monitoring of food production chains

The demand for tests that provide traceability of the taxonomic origin of biological samples or foodstuffs has increased since the crisis in the food sector triggered by the epidemiological outbreak of Bovine Spongiform Encephalopathy in the United Kingdom and the growing tendency to illegally mix meats of different taxonomic origins, without labelling the end product accordingly. For example, it has been disclosed that meat of bovine origin had been systematically added to chicken products imported to Holland from all over the world and then distributed across Europe. Other forms of adulteration have also been reported and it is believed that the practice may be widespread in industries within this sector. However, it is difficult to detect without developing advanced techniques based on DNA analysis.

Safety tests on foodstuffs are normally conducted in government laboratories, food processing plants and service companies associated with these industries. At present, users of this sector are increasingly seeking methods of control in response to growing customer demands. In this respect, certain supermarket chains have established partnerships with technological firms in order to develop methods of tracing the taxonomic origin of meat products using genetic techniques.

In a related application, the analysis of animal feed is one of the priorities of the European Agricultural Agenda, particularly after the "mad cow" crisis concerning animal feed. The analysis of foodstuffs is highly desirable and could be made compulsory in the near future. Monitoring procedures are currently based on keeping records but they do not cover the practices of illegal adulteration or dilutions that are carried out indiscriminately in many parts of the world.

Monitoring and surveillance of biodiversity

Biodiversity is the result of the interactions between the phylogenetic history of life on earth and evolutionary processes. As such, biodiversity is the sum of all life on earth and it includes genetic and functional diversity and the diversity of species.

One of the first steps in biodiversity monitoring programmes is the compilation of a taxonomic inventory, specifying all the taxa and their systematics in a specific ecosystem that includes animals, plants and microorganisms. These inventories provide the basis for all monitoring of biodiversity and conservation programmes. Global biodiversity is extremely vast: 52,629 different species of vertebrates, 4.63 million species of invertebrate and 265,876 species of plants and fungi have so far been described (figures taken from the Red List).

DNA is increasingly being accepted as the means of monitoring biodiversity. For example, DNA taken from the fur found in Canadian forests in 2002 was used to confirm that the Wild lynx still exists in the region of the great lakes.

Monitoring and surveillance of endangered species.

There are currently about 8,000 animal and plant species on official lists of endangered species, and this figure is rising each year. This trend points to the need for straightforward, "universally" usable tests for identifying the taxonomic origin of biological samples.

The trade of products made from endangered species is controlled by the, Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) and DNA tests are currently being developed under the sponsorship of this organisation to monitor this illegal trade. It is particularly important to be able to detect processed animal or plant products, such as foodstuffs or cosmetics, far more so than animal-derived raw materials such as furs, which do not require DNA testing for their identification. An example of this is the use of powdered tiger's teeth, which is widely used in traditional medicines, despite it having been declared that the survival of the species is in a grave state of danger. DNA testing has been used to identify the tiger-derived material based on the gene sequence of cytochrome b in amplified DNA and similar tests have been disclosed to trace whale meat from protected groups in processed products that apparently contain "legal" whale meat. Similar approaches have been applied to protected orchids, snakes and crocodiles in products destined for human consumption and the origin of caviar from protected sturgeons, etc. It is very likely that the use of this type of technologies for monitoring endangered species will become much more widespread in the near future.

Technology applied to the taxonomic analysis of samples

The methodology used to determine the animal origin of biological samples derives primarily from the food industry and the meat product sector. From the traditional methods based on electrophoretic and/or immunochemical analysis of proteins, technology has progressed towards the analysis of the DNA content of food samples in order to unequivocally identify the nature of the product. These methods identify nucleic acids through the hybridisation of specific probes for a specific species and/or the selective amplification of the target sequences using polymerase chain reaction (PCR).

The amplification has targeted segments of mitochondrial DNA (cfr. Bartlett et al. BioTechniques 1992 vol.12 pp.408-411; Unseld et al. Genome Research 1995 vol.4 pp.241-243; Palumbi et al. J. Hered. 1998 vol.89 pp.459-464; Wolf et al. J. Agricult and Food Chem. 1999 vol.47 pp.1350-1355; Partis et al. Meat Science 2000 vol.54 pp.369-376). This method is not suitable for determining samples with a dual or heterogeneous taxonomic origin. The amplification has also targeted nuclear DNA (cfr. Janssen et al. J. Ind. Microbiol. and Biotech. 1998 vol.21 pp.115-120, Matsunaga et al. Meat Science 1999 vol.51 pp.143-148; Wolf and Lüthy Meat Science 2001 vol.57 pp.161-168). Some of the most important proteins have been type II DNA-Topoisomerase (cfr. US 5.645.994) and α- cardiac actin (cfr. Bartlett et al. Meat Science 1998 vol.50 pp.105-114; Fairbrother et al. Animal Biotech. 1998 vol.9 pp.89-100; Lockley and Bardsley Meat Science 2002 vol.61 pp.163-168).

Some methods are based on a PCR where one oligonucleotide primer is generic and the other is dependent on the species to be identified, which is of some use in the identification of widely consumed meat species (cfr. Matsunaga et al. Meat Science 1999 vol.51 pp.143-148). Other methods have been designed to confirm the presence of DNA deriving from porcine (cfr. Montiel-Sosa et al. J. Agric. Food Chem. 2000 vol.48 pp.2829-2832), bovine, ostrich and emu species (cfr. Colombo et al. Meat Science 2000 vol.56 pp.15-17) in biological samples, but the problem arises when its presence fails to be confirmed, as this technology does not provide data on the taxonomic identity of the sample analysed.

The closest document to the present invention is US patent 5 645 994, which discloses a method for selectively amplifying DNA segments from one or more organisms in a sample through the use of gene sequences of type II DNA-Topoisomerase.

However, at present, all these methods are of limited use when the sample comprises a mixture of organisms. They would only confirm the presence of a pre-

known or suspected organism and they would not make it possible to identify each of the organisms present in the sample.

It is desirable to find a way of identifying a plurality of organisms in a single sample without having to use multiple probes and without prior knowledge of the organisms that might be present. Another aspect that could be improved is the ability to distinguish very similar or interrelated species.

DEFINITIONS

The following definitions are provided for the purposes of the present description:

Ubiquitous protein: proteins with a similar structure and function that are present in many or all of the organisms. A protein with these characteristics will be the same as the equivalent protein of the other species.

Conserved segment: Used to refer to amino acid segments and nucleotide segments. Presenting segments that are substantially or wholly common to the different species. The "high" degree of conservation indicates that the proportion of segments that several species have in common is high. This is referred to as consensus sequence.

Divergent segment: Used to refer to both amino acid segments and nucleotide segments. Presenting segments that are substantially different between different species. In this document the term "target" will also be used to refer to these sequences.

EXPLANATION OF THE INVENTION

The present invention overcomes many of the aforementioned limitations. In this regard, the inventors of the present invention have surprisingly found that the gene that codes the cytoplasmatic beta-actin protein and its derived products can be applied to taxonomic identification using samples of biological material deriving from a single species or a heterogeneous mixture of species and/or subspecies.

The present invention provides a method for identifying species and subspecies in a biological sample deriving from a single species or a heterogeneous mixture of species and/or subspecies, by means of the selective amplification of nucleic acid segments that code a target region of a macromolecule present in all the organisms concerned, for which, according to a first aspect, the object of the present invention is a method comprising a step whereby DNA is extracted from the sample; a step whereby cytoplasmatic beta-actin gene segments are amplified by

PCR or an equivalent technique; and a step whereby the amplified segment is identified by comparing its size in base pairs with a pre-established standard of sizes and/or identifying the amplified segment by DNA sequencing and comparison of the resulting sequence with the specific sequence of each species or subspecies present on a computer database.

The step whereby the starting DNA is amplified is not restricted to use of the PCR; it is possible to use any equivalent technique that can be conducted by a person skilled in the art using the tools currently available. Likewise, for example, viewing of the PCR result is not restricted to the use of electrophoresis in agarose gel; it is also possible to use capillary electrophoresis, an automated electrophoresis or any equivalent technique with a minimum resolution that is sufficient to successfully perform the experiment.

Preferably, the regions to be amplified are divergent gene segments from the cytoplasmatic beta-actin gene with DNA sequences with high evolutionary conservation between species and subspecies. And more particularly, the regions to be amplified are those which lie between the 3' sequence of the upstream exon and the 5' sequence of the downstream exon comprising the whole intronic sequence and part of the flanking exonic sequences.

In one particular embodiment of this method, the regions to be amplified are those which lie between positions 1130-1473, 1452-2063, 2438-2680 and/or 2642-2960 (numbering in relation to the DNA sequence of the human locus HUMACCYBB Accession number M10277, Genebank). In particular, the samples consist of animal tissue, more specifically horse, goat, rabbit, dog, cat, chimpanzee, human and/or brown bear tissue. In another embodiment, the samples consist of plant tissue.

In another particular embodiment of this method, in the identification step, the amplified segment or segments are compared with the human sequence M10277 and/or with the sequences of these same gene regions of species included on a computer database. The amplified segments show the conserved areas at the ends of each amplified segment and the divergence in the central region corresponding largely to the intronic region of the gene.

The present invention provides the means of identifying a plurality of organisms in a single sample without having to use multiple probes that are specific to each of the species and subspecies that might be present in the sample. The method uses universal primers, which are valid for identifying any species or subspecies present in the sample without prior knowledge of the organisms that might be present. According to the invention, a composition of universal primers are

used, which hybridise with the conserved regions of the cytoplasmatic beta-actin gene, preferably with the sequences which lie between positions 1130-1191 and 1453-1473; 1453-1473 and 2041-2065; 2433-2459 and 2643-2680 and/or 2643-2680 and 2940-2960 (numbering in relation to the DNA sequence of the human locus HUMACCYBB Accession number M10277). The particular pairs of universal primers used are P1 (1132-1151) 5'TCCGGCATGTGCAAGGCCGG3' and P2 (1474-1454) 5'CTCCATGTCGTCCCAGTTGG3'; P3 (1453-1484) 5'ACCAACTGGGACGACATGGAGAAGATCTGGC3' and P4 (2063-2034) 5'TACATGGCNGGGGTGTTAAAGGTCTCAAAC3', P5 (2434-2463) 5'TGCCCTGAGGCCCTCTTCCAGCCTTCCTTC3' and P6 (2681-2643) 5'GGGTACATGGTGGTGCCGCCAGACAGCACNGTGTTGGC3'; and P7 (2643-2681) 5'GCCAACACNGTGCTGTCTGGCGGCACCACCATGTACCC3' and P8 (2952-2932) 5'TCGTACTCCTGCTTGCTGATCCACATCTG3'.

According to a second aspect, another object of the present invention is the use of DNA sequences of the cytoplasmatic beta-actin gene in biological samples deriving from a single species or from a heterogeneous mixture of species and/or subspecies, to identify the biological species to which the samples belong.

The cytoplasmatic beta-actin protein fulfils a number of criteria for achieving reliable identification. It is a ubiquitous protein in all the organisms concerned. Cytoplasmatic beta-actin is one of the six different isoforms of actin so far identified. Specifically, cytoplasmatic beta-actin is one of the two non-muscular cytoskeletic actins. Its function is to allow mobility and provide the cell with structure and integrity, being a majority component of the cellular contractile apparatus. For this reason, it is a fundamental protein for the cell's survival, which means that it presents exonic segments with a high evolutionary conservation between species. The degree of equivalence in its amino acid secuence between species is between 98% and 100%, sufficient to present highly conserved segments but also divergent segments in the non-coding parts of the gene to correctly distinguish between species that are closely related to one another. The nucleotide divergence corresponding, for example, to intron B of the species being studied (1216-1347 bp, numbering in relation to DNA sequence of the human locus HUMACCYBB Accession number M10277) is less than 25%. Segments that are highly conserved between the different species and subspecies make it possible to use primers that are common to all the species and subspecies, whilst divergent segments are the object of amplification using said primers, resulting in a different pattern of amplification for each species and subspecies.

In addition to the qualitative identification of species present in an unknown sample, one aspect of the present invention relates to the quantitative analysis of the species present. This feature is important, for example, in determining the levels of contamination of a sample by material deriving from another species. In many cases, a qualitative result will be sufficient (for example, has chicken meat been adulterated using bovine products?), but in other cases a quantitative response will be necessary (how much of the bovine product has been added to the chicken meat?) This is particularly important when certain additives are accepted within specified limits.

Throughout the description and claims the word "comprise" and its variants do not imply the exclusion of other technical characteristics, additives, components or steps. The abstract of this application is included here by way of a reference.

For persons skilled in the art, other objects, advantages and characteristics of the invention will arise partly out of the description and partly when the invention is put into practice. The following particular embodiments and figures are provided by way of a non-limiting, illustrative example of the present invention.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 shows a diagram of the structure of the human cytoplasmatic beta-actin gene. The boxes represent the exons (exon 1 to 6) and the continuous black line represents the introns (I, Intron A to E). Regions W, X, Y and Z correspond to regions which lie between the pairs of primers P1 and P2, P3 and P4, P5 and P6, and, P7 and P8 respectively. These fragments (W, X, Y and Z) include DNA sequences that are divergent between different biological species and can be amplified using PCR using primers P1 to P8 as shown in figure 2.

Figure 2 shows the details of the oligonucleotide primers shown in figure 1. The numbering corresponds to their position on the genome sequence of the human beta-actin gene (Accession number, Genebank: M10277; Locus: HUMACCYBB). A: Adenine, C: Cytosine, G: Guanine: T: Thymine. N: position with nucleotide degeneration.

Figure 3 Top of the figure: shows the partial amino acid sequence of the cytoplasmatic beta-actin protein of three different species, Homo sapiens (man), Mus musculus (mouse) and Caenorhabditis elegans (nematode). The alignment between these sequences shows the

high degree of conservation of the cytoplasmatic beta-actin protein between species. The asterisks indicate 100% equivalence in that position between the species being compared. The numbering corresponds to the last amino acid shown according to the reference sequence in the GeneBank (refs: Hs: X00351. Mm: NM_007393.1. Ce: NM_073416.1). Middle of the figure: specifies the nucleotide sequence of the ends of exons 2 and 3 that flank intron B (W region) in said species. The exons show the nucleotide sequence in the three species being compared, divided into their corresponding codons and the amino acid residue that they code is shown below. The asterisks correspond to the nucleotide positions that are 100% conserved between the species being compared. Bottom of the figure: specifies the complete nucleotide sequence of intron B (divergent W region) in the three species being compared, to illustrate the divergence used for the identification of the species in this invention.

Figure 4 shows a diagram that illustrates the process of taxonomic identification proposed in this invention, using a biologically heterogeneous mixture. The biological sample is processed to extract the DNA and subject it to amplification by PCR. In the case that is illustrated here, the W region with primers P1 and P2 is amplified. The PCR result is viewed using standard agarose gel electrophoresis (see electrophoresis gel, left-hand lane: molecular weight marker, 100 bp ladder. Right-hand lane: bands (A and B, with approximate molecular weight expressed in base pairs bp, resulting from the PCR of the biological sample). The bands are isolated from the gel and are purified prior to undergoing DNA sequencing by standard methods. The DNA sequences obtained from each of the bands are used to interrogate a computer database that includes the sequences of the W region of biological species. The comparison of the sequences obtained using the existing sequences in the database gives the result of the identification of the species (or species) contained in the biological sample of origin.

Figure 5 shows a flow diagram illustrating the computer process for identifying the species contained in a biological sample under analysis. The DNA sequences obtained from the two fragments of the W region in the experiment shown in figure 4 are used to interrogate a database

of DNA sequences of the W region in specific species. The database shown in this case is summarised and contains 11 different species by way of an example. (Sequence 3: Cf, Canis familiaris, dog. Sequence 4: Us, Ursus species, Bear. Sequence 5: Oa, Ovis aries, goat. Sequence 6: Fc, Felis catus, cat. Sequence 7: Hs, Homo sapiens, man. Sequence 8: Ec, Equus caballus, horse. Sequence 9: Oc, Oryctolagus cuniculus, rabbit. Sequence 10: Rn, Rattus norvegicus, rat. Sequence 11: Mm, Mus musculus, mouse. Sequence 12: Dm, Drosophila melanogaster, vinegar fly. Sequence 13: Ce, Caenorhabditis elegans, nematode). The resulting comparisons with 100% equivalence, which in this case are 1:5 and 2:8, show identity with the sequences included on the database and confirm that the biological sample of origin derives from a mixture of goat and horse.

Figure 6 shows an illustration of the divergence in molecular weight and in nucleotide sequence of the W region of some of the biological species included on the database. Us, Ursus species. Oa, Ovis aries. Cf, Canis familiaris. Hs, Homo sapiens. Ec, Equus caballus. Oc, Oryctolagus cuniculus. Rn, Rattus norvegicus. Mm, Mus musculus.

Figure 7 shows an experimental example of an agarose gel electrophoresis corresponding to ten separate amplifications by PCR of the W region which lies between primers P1 and P2 of peripheral blood from eight different animal species. The numbers on each side indicate the approximate molecular weight, expressed in base pairs (bp), obtained for the W region in each of the amplifications. It is possible to observe the difference in molecular weight of this region between the animal species included. Oc: Oryctolagus cuniculus, rabbit. Cf: Canis familiaris, dog. Fc: Felis catus, cat. Us: Ursus species, Bear. Ec: Equus caballus, horse. Pt: Pan troglodytes, chimpanzee. Oa: Ovis aries, goat. Hs: Homo sapiens, man. The lanes on the left of the gel correspond to the 100 bp ladder molecular weight standard (Invitrogen). In this standard, the lowest band corresponds to 100 bp and as they ascend, each band is 100 bp greater than the one immediately below it.

DETAILED DESCRIPTION OF PARTICULAR EMBODIMENTS

<u>Identification of a species using a homogeneous / heterogeneous biological sample.</u>

The process developed for the taxonomic identification of a biologically heterogeneous sample of unknown composition is described below. The procedure would be the same for a homogeneous sample, as it is always presumed that absolutely nothing is known about the number of different species and/or subspecies present or the taxonomic nature in itself.

Processing the sample

Genome DNA was extracted from a 200 µl sample of venous whole blood in EDTA, which was compatible with any commercial kit for rapid DNA extraction, for its subsequent amplification by PCR.

The genome DNA obtained was then amplified by PCR. The W region (Figure 1) was amplified with the primers designed against nucleotide positions 1132-1151 (P1, forward primer, 5'TCCGGCATGTGCAAGGCCGG3' and 1474-1454 (P2, reverse primer, 5'CTCCATGTCGTCCCAGTTGG3'), in accordance with human sequence M10277. The PCR conditions were as follows: standard reagents, initial denaturation step at 94ºC 3 minutes followed by 35 cycles of two steps each at 94ºC 10 seconds and 68ºC 2 minutes.

First approximation: identification by molecular weight

The PCR result was viewed by standard horizontal agarose gel electrophoresis at 3% in TBE buffer. The bands that were obtained were compared with an Invitrogen 100 bp ladder-marker molecular weight standard. Figure 4 shows the results that were obtained. The comparison of the mobility of the fragments amplified in the gel using the molecular weight marker shows a molecular weight of approximately 371 and 304 base pairs. If the molecular weights of the bands obtained are compared with a database of molecular sizes obtained a priori, it is possible to make a first approximation in identifying the species present in the starting sample. Figure 7 shows a pool of ten separate amplifications by PCR of the W region that lies between primers P1 and P2 of peripheral blood from eight different animal species. It is possible to observe the difference in molecular weight of this region between the animal species included. Oc: Oryctolagus cuniculus, rabbit. Cf: Canis familiaris, dog. Fc: Felis catus, cat. Us: Ursus species, Bear. Ec: Equus caballus, horse. Pt: Pan troglodytes, chimpanzee. Oa: Ovis aries, goat. Hs: Homo sapiens, man. The left-hand lanes of the gel correspond to the 100 bp ladder molecular weight standard (Invitrogen). In a first approximation by comparison with this database of molecular weights, the bands obtained in figure 4 would correspond to goat (371 bp band) and horse (304 bp band).

Second approximation: identification by DNA sequencing

A second approximation then identifies the two bands obtained by sequencing their DNA. The bands were purified using Life Technologies' Concert Rapid PCR Purification System kit, so that their DNA could then be sequenced. The sequencing was performed cyclically in both directions with the same primers used in the initial PCR in accordance with the protocols and reagents of Applied Biosystems' ABI-Prism 310 automatic sequencing system. The two sequences that were obtained were used to interrogate a database of DNA sequences of the W region of the cytoplasmstic beta-actin gene of several species using the ClustalW program developed by the European Bioinformatic Institute of the EMBL (www.ebi.ac.uk) or an equivalent program that is available on the Internet (Figure 5). The comparisons resulted in a 100% equivalence of 1:5 and 2:8 in this case, confirming the source of the biological sample of origin, a mixture of goat and horse.